Published in final edited form as:

Biometrics. 2018 March ; 74(1): 331–341. doi:10.1111/biom.12742.

Profiling the Effects of Short Time-course Cold Ischemia on Tumor Protein Phosphorylation using a Bayesian Approach

You Wu¹, Jeremy Gaskins¹, Maiying Kong¹, and Susmita Datta²

¹Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, Kentucky, U.S.A

²Department of Biostatistics, University of Florida, Gainesville, Florida, U.S.A

Summary

Phosphorylated proteins provide insight into tumor etiology and are used as diagnostic, prognostic and therapeutic markers of complex diseases. However, pre-analytic variations, such as freezing delay after biopsy acquisition, often occur in real hospital settings and potentially lead to inaccurate results. The objective of this work is to develop statistical methodology to assess the stability of phosphorylated proteins under short-time cold ischemia. We consider a hierarchical model to determine if phosphorylation abundance of a protein at a particular phosphorylation site remains constant or not during cold ischemia. When phosphorylation levels vary across time, we estimate the direction of the changes in each protein based on the maximum overall posterior probability and on the pairwise posterior probabilities, respectively. We analyze a dataset of ovarian tumor tissues that suffered cold-ischemia shock before the proteomic profiling. Gajadhar et al. (2015) applied independent clusterings for each patient because of the high heterogeneity across patients, while our proposed model shares information allowing conclusions for the entire sample population. Using the proposed model, 15 out of 32 proteins show significant changes during one-hour cold ischemia. Through simulation studies we conclude that our proposed methodology has a higher accuracy for detecting changes compared to an order restricted inference method. Our approach provides inference on the stability of these phosphorylated proteins, which is valuable when using these proteins as biomarkers for a disease.

Keywords

Bayesian modeling; Cold ischemia; Phosphorylation; Short time-course data

1. Introduction

Protein tyrosine phosphorylation is considered to be a fundamental mechanism for regulating many cellular functions. In this specific phosphorylation, a phosphate group is added to the amino acid tyrosine on a protein. Disorders of tyrosine phosphorylation are

Correspondence to: Susmita Datta.

Supplementary Materials

A Web Appendix with additional tables and figures referenced in Section 2–5, as well as R and WinBUGS codes to implement the proposed method, is available along with this paper at the Biometrics website on Wiley Online Library.

believed to lead to many serious human diseases (Hunter, 2009). For example, protein tyrosine phosphorylation is tightly regulated in normal cells, but tyrosine kinases, whose activity controls tyrosine phosphorylation, are found to be mutated or over-expressed in many human malignancies (Paul and Mukhopadhyay, 2004). Accurate and robust assessment of tyrosine phosphorylation in tumor biopsy samples is thus necessary for understanding intracellular signaling networks and for developing targeted therapies for cancer patients (Bonnas et al., 2012; Gajadhar et al., 2015).

However, there may exist pre-analytic variations due to inconsistencies during sample collection and processing in a clinical laboratory. One of these sources of variations is cold ischemia. Also known as freezing delay time, cold ischemia is the time between tissue specimen excision and the freezing of the sample. Although it has been shown that global protein levels do not change up to one hour cold ischemia, significant changes are observed in phosphorylated proteins at the phosphorylation sites. Some of the phosphorylation sites even have rapid changes during the first 5 minutes of cold ischemia (Mertins et al., 2014; Gajadhar et al., 2015). The dynamic nature of protein phosphorylation in tissue specimens may be due to the fact that the kinases and phosphatases controlling the signaling pathways of phosphorylation are still active *ex vivo* after tissue excision (Espina et al., 2008). Therefore, the fidelity of a protein phosphorylation abundance, which is the relative intensity value used to measure protein phosphorylation magnitude, cannot be guaranteed in excised tumor tissues that undergo cold ischemia. Subsequently, the targeted therapeutic strategies and clinical decisions based on phosphorylation studies may not be accurate. To ensure trustworthy results, it is necessary to examine whether the phosphorylation level of a specific protein at a specific site is affected by short time cold ischemia. This may determine the suitability of that phosphorylated protein to be considered as a stable biomarker for a disease.

Although some phosphorylated proteins have been observed to be affected by short time cold ischemia in the literature, few of those studies comprehensively examine the stability of phosphorylation over an entire sample population (Gajadhar et al., 2015; Espina et al., 2008; Gündisch et al., 2013). As mentioned by Montana, Berk and Ebbels (2011), most metabolomic experiments only produce short time-course data with less than 10 time points, and the classical time series analysis cannot be applied. Gajadhar et al. (2015) examined the cold ischemia induced changes in phosphorylation by applying an affinity propagation clustering analysis to an ovarian tumor dataset from 5 patients. However, due to the high level of heterogeneity across patients, an independent clustering was formed for each patient, and it is challenging to draw conclusions for the entire population. In this article, we construct a novel hierarchical Bayesian model to examine whether the phosphorylation of each protein at a particular site stays stable or not under cold ischemia shock. When the phosphorylation levels at a particular site vary across time points, we further estimate the direction of the changes based on whether their abundances are increasing or decreasing between two adjacent time points. By utilizing random effects to capture the dependence between the observed phosphorylation abundances across different time points and different phosphorylated proteins, we develop a model that shares information across patients and allows us to draw conclusions for the entire sample population.

Our proposed Bayesian model is presented in Section 2. We also briefly describe an existing method proposed by Peddada et al. (2003) which can potentially be used to analyze this short time course data. The work of Peddada et al. (2003) is based on order restricted inference (ORI) for the analysis of a short time-course microarray data. In Section 3, we apply our proposed method and the competitor ORI approach to examine the stability of each phosphorylated protein under one hour cold ischemia shock using an ovarian tumor dataset from Gajadhar et al. (2015). We further verify our method using a simulation study in Section 4 and provide some concluding remarks in Section 5.

2. Models

2.1 Hierarchical Bayesian model

Suppose that there are N patients in the study. We consider J common phosphorylated proteins, and each protein has only one phosphotyrosine (pTyr) site. For each patient the abundance of phosphorylation at the j^{th} pTyr site ($j = 1, \dots, J$) is measured at T common time points. At the t^{th} time point ($t = 1, \dots, T$), A_t represents the number of minutes of freezing delay after sample excision (i.e., minutes of cold ischemia). The abundance of phosphorylation at the j^{th} pTyr site at the t^{th} time point for the i^{th} patient is denoted as Y_{ijt} ($i = 1, \dots, N; j = 1, \dots, J; t = 1, \dots, T$). The unknown true population mean of the response profile at the j^{th} pTyr site is expressed as $\mu_j = (\mu_{j1}, \mu_{j2}, \dots, \mu_{jT})'$.

Our primary goal is to examine the stability of each phosphorylated protein and then further classify the detected unstable proteins according to their changes in phosphorylation abundance between two adjacent time points. We characterize the changes for each site into $c = 2^{T-1} + 1$ classes. The null profile is set as $C_0 = \{\mu_j \in R^T: \mu_{j1} = \mu_{j2} = \dots = \mu_{jT}\}$, indicating that there is no change in phosphorylation abundance under cold ischemia shock over time. Each of the remaining classes has the form $C_r = \{\mu_j \in R^T: \mu_{j1} \square \mu_{j2} \square \dots \square \mu_{jT}\}$ ($r = 1, \dots, 2^{T-1}$), where $\square$ is either $<$ or $>$. Our primary objective is to learn if $j \in C_0$ versus any of the other classifications. The secondary objective is to learn which C_r is the most probable class for sites not in the null profile.

To model the changes of phosphorylation during one hour of cold ischemia, we consider a hierarchical random effect model of the following form:

$$Y_{ijt} = \mu_{jt} + \gamma_i + \eta_{ij} + \delta_{it} + \varepsilon_{ijt}. \quad (1)$$

The main quantity of interest μ_{jt} represents the mean abundance of phosphorylation at the j^{th} pTyr site and the t^{th} time point. Because the measurements for each patient across multiple sites and multiple time points are dependent, appropriate techniques are required to make an efficient and valid inference. The overall variation across patients is explained by γ_i ($i = 1, \dots, N$). Variation across different pTyr sites within the i^{th} patient is explained by the site-specific random effect $\eta_i = (\eta_{i1}, \eta_{i2}, \dots, \eta_{iT})'$. The patient-specific temporal effect is captured by the random effect $\delta_i = (\delta_{i1}, \delta_{i2}, \dots, \delta_{iT})'$, representing the time effect for the i^{th} subject. The random error ε_{ijt} is assumed to be normally distributed with mean 0 and variance σ^2 .

In this work, a Bayesian approach is considered for the inference. Due to normality of the random errors, the data is distributed as $Y_{ijt} \sim \mathcal{N}(\mu_{jt} + \gamma_i + \eta_{ij} + \delta_{it}, \sigma^2)$. The overall effect from the i^{th} patient, γ_i is modeled by a normal distribution with mean 0 and variance σ_γ^2 .

The site-specific random effect within the i^{th} patient is modeled by $\eta_{ij} \sim \mathcal{N}(0, \sigma_\eta^2)$ ($j = 1, \dots, J$).

The distribution of the temporal effects is $\delta_{it} \sim \text{MVN}(0, \sigma_\delta^2 R(\rho))$. We assume an auto-regressive correlation structure for $R(\rho)$ with $\text{corr}(\delta_{it}, \delta_{it'}) = \rho^{|A_t - A_{t'}|/s}$, although other choices are possible. Here, ρ represents the correlation between responses s minutes apart, and we use the value of $s = 10$ minutes throughout. Recall that A_t represents the number of minutes of cold ischemia.

Our primary goal is to determine if there are overall changes in the abundance at the j^{th} site, that is, if $j \in C_0$ or if j is in one of the other profiles. To that end, we introduce the random variable $Z_j \sim \text{Bernoulli}(\theta)$ to indicate whether there is variation in phosphorylation abundance at the j^{th} pTyr site during cold ischemia. The parameter θ represents the overall proportion of sites with ischemia-induced changes. The prior distribution of the trajectory μ_{jt} is set as $\mu_{jt} | (\mu_j^*, Z_j) \sim \mathcal{N}(\mu_j^*, Z_j \sigma_\mu^2)$, where $\mu_j^* \sim \mathcal{N}(\mu_0, \sigma_{\mu^*}^2)$ gives the average value across all the T time points. If $Z_j = 0$, the μ_{jt} s are equal to μ_j^* across all the time points, i.e., the j^{th} pTyr site is classified to the null profile C_0 . If $Z_j = 1$, μ_{jt} varies across time, and the j^{th} site is impacted by cold ischemia. A site with $Z_j = 1$ is classified to one of the candidate profiles $C_1, \dots, C_{c-1}$ based on the inequality directions in μ_{jt} across all the time points. The prior distribution of the overall mean μ_0 is set as normal distribution with mean 0 and standard deviation 100.

Conjugate prior distributions for the error variance σ^2 and fixed effect variance components σ_μ^2 and $\sigma_{\mu^*}^2$ are set as inverse-gamma distribution with both shape and scale parameters at

0.1. Since we expect that the overall patient effect γ_i , patient-specific temporal effect δ_{it} and site-specific effect η_{ij} have smaller impact on the response compared to μ_{jt} , the prior distributions for the standard deviations σ_γ , σ_η and σ_δ are set as half-Cauchy distributions with medians (equivalent to the scale parameter) of 0.1. The half-Cauchy is a common prior for the standard deviation of a regression coefficient or shrinkage effect as it has a non-zero density at 0 (unlike inverse-gamma distribution) allowing a small value for the variance term. Further, the heavy tail of the half-Cauchy prior protects from over-shrinkage by allowing large values when the effect is highly influential (Gelman, 2006; Carvalho, Polson and Scott, 2010). The correlation parameter ρ is given a uniform distribution from 0 to 1 as a prior distribution. The prior for θ , the population proportion of sites impacted by cold ischemia, is taken to be $\text{Beta}(1, 1)$ (i.e., Uniform[0, 1]).

2.2 Estimation and inference of profile classification

Markov chain Monte Carlo (MCMC) sampling is applied to obtain a posterior sample of size M . For each phosphorylation site j , denote the m^{th} sample of the posterior trajectory μ_{jt} as $\mu_j^{(m)}$ ($m = 1, \dots, M$). First, we consider testing $H_{0j}: Z_j = 0$ vs. $H_{1j}: Z_j = 1$ to examine whether the j^{th} site is impacted by cold ischemia. $P(Z_j = 1 | Y)$ is estimated by the proportion

of MCMC iterations with $Z_j^{(m)} = 1$. The j^{th} site is classified to the alternative profiles if $\hat{P}(Z_j = 1 | Y)$ is greater than 0.5, which is the Bayes decision rule corresponding to the 0 – 1 loss function $L(Z_j, \hat{Z}_j) = I(Z_j \neq \hat{Z}_j)$. If we choose to penalize Type I and Type II errors differently, the 0.5 threshold can be adjusted. Placing a non-degenerate prior on θ , the overall probability of changes, guarantees an automatic multiplicity correction over our tests based on Z_j (Scott and Berger, 2010). This approach allows us to maintain the false discovery rate when we consider whether μ_j varies across T time points (i.e., $Z_j = 1$) over a potentially large number of sites J .

If $H_{0j}: Z_j = 0$ is rejected, our secondary objective is to estimate the changes in direction of the unstable phosphorylated proteins. To estimate the profile of the j^{th} site, we consider two competing estimators: the maximum a posteriori (MAP) estimator and maximum pairwise (MPW) estimator. Both estimators are formed conditionally on the conclusions for H_{0j} vs. H_{1j} . If the posterior probability $P(Z_j = 0 | Y)$ is larger than the threshold value (typically, 0.5), then the j^{th} site maintains a membership in the null profile C_0 , indicating no change in phosphorylation abundance during cold ischemia. Otherwise, we reject H_{0j} , conclude that the j^{th} pTyr site varies across T time points, and classify the j^{th} site to its most likely profile. Using the MAP estimator, we place the j^{th} pTyr site that rejects H_{0j} to the profile that maximizes the estimated posterior probability $\hat{P}(\mu_j \in C_r | Y)$ ($r = 1, \dots, c - 1$), which is the profile with the majority votes among the $MMCMC$ iterations, $\arg \max_r \sum_m I(\mu_j^{(m)} \in C_r)$.

For the MPW estimator, the pTyr site that rejects H_{0j} is classified to one of the candidate profiles ($C_1, \dots, C_{c-1}$) by estimating the direction of change between adjacent measurements pairwise. Choosing the inequality between a pair of values $\mu_{j,t-1}$ and $\mu_{j,t}$ is based on choosing the larger of $\hat{P}(\mu_{j,t-1} < \mu_{j,t} | Y)$ and $\hat{P}(\mu_{j,t-1} > \mu_{j,t} | Y)$. The sequence of pairwise-estimated inequalities is used to determine the class of the phosphorylated protein.

In general, one would expect that the MAP estimator produces more accurate classification as it takes the full vector μ_j into consideration. However, the MAP estimator is based on the number of iterations that μ_j visits each of the 2^{T-1} non-null profiles C_r during the MCMC run. When T is large relative to the length of the MCMC chain, even the most probable profiles are visited infrequently, and the estimated posterior probabilities $\hat{P}(\mu_j \in C_r | Y)$ may have large variabilities relative to the differences between the more likely profiles. Conversely, the marginal pairwise probabilities $\hat{P}(\mu_{j,t-1} < \mu_{j,t} | Y)$ and $\hat{P}(\mu_{j,t-1} > \mu_{j,t} | Y)$ typically have small variabilities. Therefore, the MPW estimator may provide preferable classification for a large T . This issue is similar in spirit to choosing between the MAP and median model estimators in Bayesian variable selection (see e.g., George and McCulloch, 1993; Barbieri and Berger, 2004).

To describe the confidence in classification, we form a credible set $C_{j,1-\alpha}$ for each site j where $\hat{P}(\mu_j \in C_{j,1-\alpha} | Y) = 1 - \alpha$. For the MAP estimator, the profiles that are included in the $100(1 - \alpha)\%$ credible set are determined by their estimated posterior probabilities. For each phosphorylated protein, we sort $\hat{P}(\mu_j \in C_0 | Y)$, $\hat{P}(\mu_j \in C_1 | Y)$, ..., $\hat{P}(\mu_j \in C_{c-1} | Y)$ in decreasing order. Then the top candidate profiles are included in the credible set until their cumulative posterior probability is greater than $100(1 - \alpha)\%$. For the MPW estimator, a similar

procedure can be performed using the product of the posterior probabilities between each of the two adjacent time points to approximate the posterior probability of each profile.

2.3 ORI method

As a comparison to our proposed methodology, we also apply an order restricted inference (ORI) algorithm developed by Peddada et al. (2003). This ORI method is one of the most commonly used methods developed for analyzing short time-course microarray data. It can also be applied to this phosphorylation abundance data. We briefly describe the method here. The sample mean vector of the j^{th} pTyr site is denoted as $\bar{Y}_j = (\bar{Y}_{j1}, \dots, \bar{Y}_{jT})'$.

First, a collection of the candidate profiles $C_1, \dots, C_p$ is specified. The collection may or may not be the full set of 2^{T-1} possible non-zero profiles. Second, for each candidate profile, point estimation follows the procedures proposed by Hwang and Peddada (1994), which use an isotonic regression estimator for the general ordering cases. For each profile, the nodal parameter is the key for the estimation procedure. The nodal parameters in each profile are defined as the parameters which are linked to all the other parameters, where the two parameters are linked if the inequalities between them are pre-specified. For example, in the case that $T = 4$, nodal parameters are μ_1, μ_2, μ_3 and μ_4 in the candidate profile $C = \{\mu_1 < \mu_2 < \mu_3 < \mu_4\}$; in the candidate profile $C = \{\mu_1 < \mu_2 < \mu_3 > \mu_4\}$, μ_3 is the only nodal parameter (see Table A1 in the Web Appendix for the details when $T = 4$). For the nodal parameters, an ordered sequence of parameters can be formed as $\mu_{j(1)} \mu_{j(2)} \dots \mu_{j(T)}$, where one may arbitrarily assign an inequality between the two parameters if their relationship is unknown. Based on the ordered sequence, the estimate of the nodal parameter μ_{ji} is found using the following formula:

$$\hat{\mu}_{ji} = \hat{\mu}_{j(s)} = \min_{r \geq s} \max_{q \leq s} \frac{\sum_{k=r}^q n_{j(k)} \bar{Y}_{j(k)}}{\sum_{k=r}^q n_{j(k)}},$$

where $\bar{Y}_{j(k)}$ is the sample mean based on $n_{j(k)}$ observations. For a non-nodal parameter, say μ_{jt} , the largest sub-profile with μ_{jt} as a nodal parameter needs to be identified, then μ_{jt} is estimated by using the formula above for the nodal parameters and the data corresponding to the sub-profile. For the parameters in a profile without any nodal parameters, the largest sub-profile with at least one nodal parameter needs to be identified, and the parameters in the sub-profile can be estimated using the methods for the profiles with nodal parameters. The procedure is repeated until all of the parameters are estimated. Once a parameter is estimated, the estimate of this parameter should be used in the estimation of the other unknown parameters.

Hypothesis testing for $H_0: \mu_j \in C_0$ vs. $H_1: \mu_j \in \cup_{k=1}^p C_k$ (equivalent to our test of $H_0: Z_j = 0$ vs. $H_1: Z_j = 1$) is performed using a bootstrap approach. In each bootstrap step, the maximum difference of two linked parameters is calculated for each candidate profile. Then the largest maximum difference among the p candidate profiles can be obtained for each bootstrap sample. The hypothesis test is performed based on the bootstrap distribution of the

largest maximum difference. If H_0 is rejected, the site is classified to the profile with the largest maximum difference. Full details can be found in Peddada et al. (2003).

3. Application to an ovarian tumor dataset

3.1 The ovarian tumor data

The ovarian tumor data from Gajadhar et al. (2015) is used to illustrate our proposed methodology. Ovarian tumor tissues were collected from $N = 5$ patients. Tumor specimens were collected across $T = 4$ time points at 0, 5, 30 and 60 minutes of freezing delay after the surgical removal. After protein extraction, digestion and phosphotyrosine peptides enrichment, the peptides are separated using a mass-spectrometry based analysis. The phosphotyrosine peptides and proteins are identified and quantified subsequently.

Only the phosphorylated proteins with one phosphotyrosine position are considered in the study. The values of phosphorylation abundance are relative intensities determined by iTRAQ quantification. For relative quantification of the four time points, peptide samples were chemically labeled with different iTRAQ reagents and then mixed into a single sample. These tags get covalently attached to the amine groups of the N termini of peptides and lysine side chains. The mixed sample then get fragmented in the mass spectrometer and reporter ions at defined masses (including the masses of the pTyr sites) get released. Ratios of these reporter ions to one another are treated as relative abundance, and the intensity of the masses at the initial time point is treated as the reference abundance (Gajadhar et al., 2015). In order to find the general characteristics of the phosphorylated sites, only the phosphorylated proteins that are common to all patients are included in the following analysis. After excluding the proteins with multiple phosphotyrosine positions, $J = 32$ common pTyr sites are used in the analysis.

3.2 Results

The sampling procedure for our Bayesian methods is implemented using WinBUGS (Spiegelhalter et al., 1999) and R (R Core Team, 2014). Copies of the WinBUGS and R codes can be found in the online supplementary materials. Running the MCMC sampling procedure for 50,000 iterations with a burn-in of 5000 and thinning rate of 5 takes 3.02 minutes. The hypothesis $H_{0j}: Z_j = 0$ is rejected if $\hat{P}(Z_j = 1 | Y) > 0.5$, and both MAP and MPW estimators are used to classify each phosphorylated protein. The testing and classification results based on these two Bayesian estimators are exactly the same.

Figure 1 shows the estimated trajectories with 95% credible intervals (CI) (1st and 3rd columns) and MPW estimators (2nd and 4th columns) of six representative phosphorylated proteins. The gray lines represent the observed phosphorylation abundance of each patient. The boxes in the MPW plots represent the posterior probabilities of an equal, decreasing, or increasing trend of phosphorylation abundance between each pair of adjacent time points, and the shaded box indicates the MPW-chosen classification. For example, both phosphorylation sites of DYRK1A and CDK1 are classified to the null profile C_0 with $\hat{P}(Z = 0 | Y) = 0.98$ and 0.95, respectively. Sites of FYB, MAPK7, GAB1 and MAPK14 are classified to the non-null profiles.

Figure 1 also shows that some phosphorylation sites have similar phosphorylation trajectories across all five patients (DYRK1A and MAPK7), while some other sites (CDK1 and MAPK14) have greater heterogeneities across patients. For example, the flat trajectory seen in phosphorylated protein CDK1 may be partly due to the conflicting effects of the patient-dependent variations. The plots of estimated profiles and MPW estimator for all phosphorylated proteins can be found in Figures A1–A4 in the Web Appendix.

Detailed classification results from the Bayesian model are shown in Table 1 with the probabilities of being classified to the null profile (4th column) and the posterior probabilities of being classified to the corresponding profiles based the MAP estimator (5th column). The profiles included in the 75% credible set are reported in the last column in Table 1. Recall the site's trajectory is estimated to be one of the profiles in this set with at least 75% posterior probability. 17 pTyr sites were classified to the null profile, indicating that the phosphorylation abundance does not change during one-hour cold ischemia based on both estimators. The pTyr sites of DOK1, INPP5D and MPZL1 show an increasing trend (C_1), while the sites of CDKL5, EPHA2, FYB and PRKCD show a decreasing trend during one hour cold ischemia (C_8). Some of our classification results are consistent with the clustering results in Gajadhar et al. (2015). For example, ANXA2 and FLNA were classified to the null profile C_0 based on our proposed model, and they were also classified to the cluster with minimal quantitative fluctuations in Gajadhar et al. (2015). MAPK14 showed a rapid increase in phosphorylation within 5 minutes cold ischemia in Gajadhar et al. (2015), and it was classified to the candidate profile C_3 with similar characterization based on our Bayesian model. However, the conclusions in Gajadhar et al. (2015) are based on a single sample, while our conclusions are based on all samples and should be more robust.

Table 2 shows the posterior median and 95% credible interval (CI) for each model parameter. The estimated standard deviations for the patient and site-specific components, σ_γ and σ_η , are quite small, indicating the impact of our half-Cauchy shrinkage prior. Conversely, the standard deviation of the temporal effect σ_δ is much larger (roughly 60% of the standard deviation for error), while the correlation is low ($\rho = 0.13$). Thus, variation across T time points is the key driver of the covariance structure, not the variation across patients or sites. By including the within patient temporal effect and site-specific effect, we are able to cohesively distinguish between the roles of protein trajectory μ_{jt} and the patient trajectory δ_{jt} , while Gajadhar et al. (2015) performs patient-specific analysis which potentially confounds the two sources of variation.

Classification was also done using the ORI method. Detailed results of ORI profiling with p-values for the test of $H_0: \mu_j \in C_0$ versus $H_1: \mu_j \in \cup_{k=1}^{c-1} C_k$ can be found in Table A2 in the Web Appendix. Table A3 in the Web Appendix compares the classifications from the Bayesian approach to those from the ORI method. Of the 17 pTyr sites classified to the null profile by the Bayesian method, the ORI method classifies 14 of these to the null profile and 3 to the non-null profiles. There is one profile which the ORI method classifies to the null profile, however, the Bayesian method classifies to the non-null profile. In total, 23 pTyr sites (71%) were classified to the same profiles using both methods. The total computation

time using the ORI method was 26.4 minutes, almost seven times longer than our Bayesian method.

4. Simulation study

To verify our approach and compare the results with the ORI method, we conduct a simulation study with varying sample sizes and number of time points. In each simulation, N subjects are generated, and each subject is observed at T time points and J pTyr sites. We assign the true profile for each of the J pTyr sites as follows. When $T = 4$, the 20% of sites are assigned to the null profile C_0 , then 5% to profile C_1 , and similarly, 10%, 10%, 5%, 15%, 5%, 15% and 15% of sites are assigned to profiles C_2 to C_8 , respectively. The trajectories μ_{jt} s are randomly sampled from the sequence from -3 to 5 with increments 0.5 according to these pre-specified profiles. Data are generated based on model (1) with the error variance σ^2 set to 0.45 . The variance components σ_η^2 , σ_γ^2 , and σ_δ^2 are all set as 0.1 . The temporal correlation coefficient ρ is set as 0.5 . 200 simulated datasets are generated and analyzed using both ORI algorithm and our Bayesian model.

4.1 Assessing the effect of sample size

The first simulation study is run with $N = 5$, $J = 35$, $T = 4$, and the time points are chosen as 0, 5, 30, and 60 minutes, which is similar to the setting of ovarian tumor data. Another simulation study is carried out with an increased sample size of $N = 20$. We draw 200 simulated datasets for each simulation setting. To evaluate the performance of our procedure, we consider the following four quantities: Type I error rate, Type II error rate, the overall correct classification proportion, and the accuracy of trajectory estimation. To measure the accuracy of the estimation for μ , we use a sum of squared error (SSE) loss function $L(\mu, \hat{\mu}) = \sum_{j=1}^J \sum_{t=1}^T (\hat{\mu}_{jt} - \mu_{jt})^2$, based on the posterior mean of $\hat{\mu}_{jt}$. All accuracy measurements are averaged over 200 simulated datasets.

Table 3 shows the simulation results of the two scenarios introduced above with $N = 5$ and $N = 20$. For the simulation with $N = 5$, the overall mean proportion of correct classification using the ORI method is 68.3%, compared to 94.1% from our Bayesian model using both the MAP and MPW estimators, corresponding to roughly 9 additional correctly assigned sites. The probabilities of making a Type I error $P(\mu_j \in \cup_{k=1}^{c-1} \hat{C}_k | \mu_j \in C_0)$ in our proposed model and ORI methods are 4% and 5%, respectively. The Type II error rates $P(\mu_j \in \hat{C}_0 | \mu_j \in \cup_{k=1}^{c-1} C_k)$ are around 0.2% for both the MAP and MPW estimators, compared to 0.4% based on the ORI method, indicating that all methods have high power in this scenario. The measurement for the accuracy of the estimated trajectory (lower SSE corresponds to a greater accuracy) is 18.5 for the Bayesian model compared to 54.1 for the ORI method, indicating that our model more accurately estimates the trajectory μ_j than the ORI method.

For the simulation with increased sample size with $N = 20$, the mean of the correct classification proportions of pTyr sites using the ORI method is 71.2% compared to 99.5% using both the estimators of our Bayesian model. Both of our proposed model and the ORI

method accurately assign sites to the null, while our Bayesian model does better in distinguishing the non-null profiles. In this scenario, our Bayesian model also performs better for estimating the trajectory than the ORI method.

We also construct the 75% credible sets in both the simulation settings. The proportions that the true profile is included in the 75% credible set are 98.5% and 99.9% for $N=5$ and 20, respectively. Thus, we have confidence that in the rare case when the estimated profile is incorrect, the true classification is almost always in the credible set.

4.2 Assessing the effect of the signal-to-noise ratio and number of time points

We conduct additional simulation studies to examine the performance of our model under different signal-to-noise ratios and different numbers of time points. In the following simulation scenarios, we let $N=20$ and $J=35$ and consider three settings for T : $T=4, 5$, and 8. For $T=4$, each site is assigned to one of the 9 candidate profiles as in the previous section. However, as T increases, the number of candidate profiles increases exponentially. This is particularly challenging for the ORI method because its algorithm involves a recursive estimation scheme for each profile under consideration. To that end, we apply the ORI method to a restricted set of profiles, which are known to contain the set of true profiles. For $T=5$, 11 out of 17 candidate profiles (including the null) are selected, representing 65% of the total number of possible profiles. For $T=8$, the restricted set contains 15 out of 129 profiles, roughly 12% of all possible profiles. Table A4 in the Web Appendix reports the selected profiles and the percentage of the true assignment for the cases when $T=5$ and 8.

In addition to examining the impact of T on accuracy, we also consider changes to the signal-to-noise ratio (i.e., effect size). For the largest effect size, μ_{jt} are drawn uniformly from the sequence from -3 to 5 with an increment of 0.1 , subject to the profile membership. The medium effect size draws μ_{jt} from the sequence from -1.5 to 2.5 with an increment of 0.05 , and the smallest effect size uses -0.75 to 1.25 with an increment of 0.025 . The ORI method is applied to find the best profile among the selected set of the candidate profiles. To determine the accuracy without any additional knowledge of the restricted profile set, the MAP and MPW classifications are found as before. To compare against the ORI estimator that uses the selected profile set, we develop corresponding versions of the MAP and MPW estimators that use this extra knowledge. For the MAP estimator, we choose the candidate profile which has the maximum among $Pr(\mu_j \in C_r | Y)$ among r in the selected profile set. For the MPW estimator, we choose the profile from among the selected set with the maximum value of $\prod_{t=1}^{T-1} Pr(\mu_{jt} \square_t \mu_{j,t+1} | Y)$, where $\square_t$ is the t^{th} inequality in C_r .

Figure 2 shows the boxplots of the correct classification rates under the different simulation scenarios. In the $T=4$ case, the mean correct classification rates of both the MAP and MPW methods are higher than the ORI method under all effect sizes. As the number of time points increases and/or the effect size decreases, correct classification becomes more challenging, and the accuracy deteriorates for all methods. However, across all scenarios the Bayesian methods that use the selected set of profiles consistently beat (often by a large margin) the ORI in accuracy. In fact, even without using the extra information on the restricted set of

profiles, our Bayesian classification is often more accurate than the ORI (e.g., for all the effect sizes for $T=4$ and 5, and for the large effect size for $T=8$). Comparing the two estimators, we find that the MAP and MPW estimators are roughly equivalent for the small number of time points, and the MPW estimator is slightly better for the larger number of time points, which is consistent with the intuition discussed in Section 2.2. See Table A5 in the Web Appendix for detailed results of each simulation scenario. Figure 3 shows the boxplots of the accuracy of the estimated trajectories measured by SSE. In all cases, the Bayesian model performs better than the ORI method in estimating the trajectory μ_{jt} .

5. Discussion and Conclusions

In this project we develop a Bayesian methodology to examine the stability of tyrosine phosphorylated proteins undergoing cold ischemia and to characterize the direction of changes in the unstable ones. From the ovarian tumor data, 15 out of 32 phosphorylated proteins show significant changes during one-hour cold ischemia based on our Bayesian model, indicating that any scientific results related to these phosphorylated proteins may be altered by a short freezing delay after the tumor specimen excision.

A number of phosphorylated proteins that we identified as fluctuating in phosphorylation abundance under cold ischemia shock have been found to be medically important for therapy development or biomarker discovery in the literature. For example, DOK1 has been identified as a candidate tumor suppressor gene for various human malignancies (Lee et al., 2007; Berger et al., 2010) and works in a signal transduction pathway downstream of receptor tyrosine kinases (Némorin et al., 2001). This phosphorylated protein shows an increasing trend in phosphorylation abundances undergoing one hour cold ischemia shock in our study. GAB1, which serves in a different signaling pathway, was classified to the profile with an increase in phosphorylation abundance during the first 30 minute cold ischemia and a decrease during the next 30 minute cold ischemia based on our Bayesian model; it was also classified to the cluster characterized by a rapid increase in phosphorylation within 5 minutes of cold ischemia in Gajadhar et al. (2015). It has been found that GAB1 is usually over-expressed in cancer cells and may be used as a target for cancer therapy, especially for triple negative breast cancer patients (Chen et al., 2015; Gu and Neel, 2003). Similarly, EPHA2 has been found to be over-expressed in various cancers, particularly a high level of EPHA2 is detected in malignant cancer-derived cell lines and advanced forms of cancer (Tandon, Vemula and Mittal, 2011). EPHA2 displayed a decreasing trend in phosphorylation abundance during one hour cold ischemia based on our approach and was also classified to the gradually decreased cluster in Gajadhar et al. (2015). It should be noted that PEAK1 may require a more careful sample collection protocol compared to the other phosphorylated proteins classified to the null profile. While the most probable explanation is that PEAK1 is not impacted by freezing delay, the non-null profile $C_6 = \{\mu \in R^4 : \mu_1 > \mu_2 < \mu_3 > \mu_4\}$ is contained in the credible set, so the potential impact of cold ischemia should not be totally discounted.

Based on the simulation studies, our Bayesian model is more efficient for the classification of short time-course data than the ORI algorithm. Both our Bayesian model and the ORI method have high correct classification rates for the null profile C_0 . However, our Bayesian

model has a much higher accuracy in correctly classifying the non-null profiles than the ORI method. The Bayesian approach performs well even in the cases with more time points and small effect sizes. In addition to the improved accuracy, the computation speed of our Bayesian approach is much faster than the ORI method. In the simulation study with $T = 8$, it takes approximately 6 hours per dataset to use the ORI method on the subset of 15/129 profiles, whereas the Bayesian approach which considers all profiles requires only a third of the time.

Since the primary focus is whether the protein phosphorylation status has changed during one-hour cold ischemia, we only consider complete equality or a sequence of strict inequalities between each pair of time points. If the changes in phosphorylation abundance under a certain threshold need to be considered as equal, some extensions of our model can be considered. Wu et al. (2007) provides a similar approach using a 3-stage Markov model using the relationships $<$, $>$ and $=$ between two time points. However, they do not consider the role of the site or temporal dependence. As a simple ad hoc version, one might extend our method to allow the profiles with relationships $<$, $>$ and $=$ by choosing a cut-off value for approximate equality either through biological considerations or a default value. The pair μ_{jt} and $\mu_{j,t+1}$ would be considered to be approximately equal if $|\mu_{jt} - \mu_{j,t+1}| < \epsilon$, increasing if $\mu_{jt} + \epsilon < \mu_{j,t+1}$, and decreasing if $\mu_{jt} - \epsilon > \mu_{j,t+1}$, where ϵ is a pre-specified positive value. A MAP or MPW estimator from the set of 3^{T-1} profiles can be found.

A key concern for an analysis with few patients is the validity of the normality assumption. A Q-Q plot and histogram of the standardized residuals $(Y_{ijt} - \hat{\mu}_{jt} - \hat{\gamma}_i - \hat{\delta}_{it} - \hat{\eta}_{ij})$ can be found in Web Appendix, Figure A5. The Q-Q plot indicates that the normality assumption is reasonably satisfied in our analysis of the ovarian tumor data. If normality was found to be suspect, a transformation of the data (e.g., logarithm or Box-Cox) or a thicker-tailed or skewed distribution for ϵ_{ijt} , such as the t-distribution or skewed normal distribution, could easily be incorporated.

The phosphotyrosine signaling networks can be affected by cold ischemia shock for many phosphorylated proteins, even in a short time period. Our study shows that nearly half of the selected phosphorylated proteins experience a dramatic change during one-hour cold ischemia, which underscores the necessity of freezing the tissue sample immediately after excision. Because the phosphorylation changes of some pTyr sites classified to C_0 may be dramatically different among patients (see MAPK14 in Figure 1), these sites may require further attention. The current work provides an efficient method to detect unstable phosphorylated proteins under short-time cold ischemia shock and profile the direction of their changes. Thus, the results obtained from our method may provide valuable guidance on developing sample collection protocols and further analytical strategies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) grant R15CA170091.

References

- Barbieri MM, Berger JO. Optimal predictive model selection. *Annals of Statistics*. 2004; 32:870–897.
- Berger AH, Niki M, Morotti A, Taylor BS, Socci ND, Viale A, Brennan C, Szoke J, Motoi N, Rothman PB, Teruya-Feldstein J. Identification of DOK genes as lung tumor suppressors. *Nature Genetics*. 2010; 42:216–223. [PubMed: 20139980]
- Bonnas C, Specht K, Spleiss O, Froehner S, Dietmann G, Krger JM, Arbogast S, Feuerhake F. Effects of cold ischemia and inflammatory tumor microenvironment on detection of PI3K/AKT and MAPK pathway activation patterns in clinical cancer samples. *International Journal of Cancer*. 2012; 131:1621–1632. [PubMed: 22213219]
- Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. *Biometrika*. 2010; 97:465–480.
- Chen L, Du-Cuny L, Moses S, Dumas S, Song Z, Rezaeian AH, Lin HK, Meuillet EJ, Zhang S. Novel inhibitors induce large conformational changes of GAB1 pleckstrin homology domain and kill breast cancer cells. *PLOS Computational Biology*. 2015; 11:e1004021. [PubMed: 25569504]
- Espina V, Edmiston KH, Heiby M, Pierobon M, Sciro M, Merritt B, Banks S, Deng J, VanMeter AJ, Geho DH, Pastore L. A portrait of tissue phosphoprotein stability in the clinical tissue procurement process. *Molecular & cellular Proteomics*. 2008; 7:1998–2018. [PubMed: 18667411]
- Gajadhar AS, Johnson H, Slebos RJ, Shaddox K, Wiles K, Washington MK, Herline AJ, Levine DA, Liebler DC, White FM. Phosphotyrosine signaling analysis in human tumors is confounded by systemic ischemia-driven artifacts and intra-specimen heterogeneity. *Cancer Research*. 2015; 75:1495–1503. [PubMed: 25670172]
- George EI, McCulloch RE. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*. 1993; 88:881–889.
- Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*. 2006; 1:515–534.
- Gu H, Neel BG. The Gabin signal transduction. *Trends in Cell Biology*. 2003; 13:122–130. [PubMed: 12628344]
- Gündisch S, Grundner-Culemann K, Wolff C, Schott C, Reischauer B, Machatti M, Groelz D, Schaab C, Tebbe A, Becker KF. Delayed times to tissue fixation result in unpredictable global phosphoproteome changes. *Journal of Proteome Research*. 2013; 12:4424–4434. [PubMed: 23984901]
- Hunter T. Tyrosine phosphorylation: thirty years and counting. *Current Opinion in Cell Biology*. 2009; 21:140–146. [PubMed: 19269802]
- Hwang JG, Peddada SD. Confidence interval estimation subject to order restrictions. *The Annals of Statistics*. 1994; 22:67–93.
- Lee S, Huang H, Niu Y, Tommasino M, Lenoir G, Sylla BS. Dok1 expression and mutation in Burkitt's lymphoma cell lines. *Cancer Letters*. 2007; 245:44–50. [PubMed: 16338067]
- Mertins P, Yang F, Liu T, Mani DR, Petyuk VA, Gillette MA, Clauser KR, Qiao JW, Gritsenko MA, Moore RJ, Levine DA. Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Molecular & Cellular Proteomics*. 2014; 13:1690–1704. [PubMed: 24719451]
- Montana, G., Berk, M., Ebbels, T. *Software Tools and Algorithms for Biological Systems*. Springer; New York: 2011. Modelling short time series in metabolomics: a functional data analysis approach; p. 307-315.
- Némorin JG, Laporte P, Bérubé G, Duplay P. p62dok negatively regulates CD2 signaling in Jurkat cells. *The Journal of Immunology*. 2001; 166:4408–4415. [PubMed: 11254695]
- Paul MK, Mukhopadhyay AK. Tyrosine kinase-role and significance in cancer. *International Journal of Medical Sciences*. 2004; 1:101–115. [PubMed: 15912202]
- Peddada SD, Lobenhofer EK, Li L, Afshari CA, Weinberg CR, Umbach DM. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics*. 2003; 19:834–841. [PubMed: 12724293]
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2014. URL <http://www.R-project.org/>

- Scott JG, Berger JO. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*. 2010; 38:2587–2619.
- Spiegelhalter DJ, Thomas A, Best NG. WinBUGS Version 1.2 User Manual. MRC Biostatistics Unit. 1999
- Tandon M, Vemula SV, Mittal SK. Emerging strategies for EphA2 receptor targeting for cancer therapeutics. *Expert Opinion on Therapeutic Targets*. 2011; 15:31–51. [PubMed: 21142802]
- Wu H, Yuan M, Kaech SM, Halloran ME. A statistical analysis of memory CD8 T cell differentiation: An application of a hierarchical state space model to a short time course microarray experiment. *The Annals of Applied Statistics*. 2007; 1:442–458.

Figure 1. Estimated trajectory with 95% credible interval (1st and 3rd columns) and MPW estimators (2nd and 4th columns) for six representative phosphorylated proteins. The gray lines represent the observed phosphorylation abundance of each of the five patients. The boxes of MPW plots represent the posterior probabilities between each of two adjacent time points, and the shaded box indicates the MPW-selected classifications.

Figure 2.

Boxplots of simulation results from Section 4.2. The correct classification proportions shown are calculated based on 200 simulated datasets for each case. For $T=4$, all the possible candidate profiles are considered. For $T=5$ and 8, the ORI method is not used when considering all possible profiles due to the limitation of computation speed.

Figure 3. Boxplots of sum of squared error (SSE) measuring the accuracy of trajectory estimation. A lower value indicates higher accuracy.

Table 1

Detailed Result of Bayesian classification based on MAP. The probability of each protein being classified to the null profile is reported in 4th column with the column title “ $\hat{P}(Z_j = 0|Y)$ ”. The posterior probabilities of classifying each protein to its corresponding profile is reported in the 5th column with the column title “Maximum Posterior Probabilities”. The candidate profiles included in 75% credible set for each protein are reported in the last column.

Protein	pTyr	Sequence	$\hat{P}(Z_j = 0 Y)$	Maximum Posterior Probabilities	75% Credible Set
$C_0 = \{\mu \in R^4 : \mu_1 = \mu_2 = \mu_3 = \mu_4\}$					
annexin A2 isoform 2	ANXA2 *	LSLEGDHSTPPSA[y]GSVK	0.970	0.970	C_0
glucocorticoid receptor DNA binding factor 1	ARHGAP35 *	NEEEN[y]SVPHDSTQGK	0.979	0.979	C_0
cell division cycle 2 protein isoform 1	CDK1 *	IGEGT[y]GVVYK	0.951	0.951	C_0
dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1A isoform 3	DYRK1A	IYQ[y]IQSR	0.984	0.984	C_0
dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 2 isoform 1	DYRK2 *	VYT[y]IQSR	0.974	0.974	C_0
ephrin receptor EphB3 precursor	EPHB3 *	FLEDDPSDPT[y]TSSSLGGK	0.850	0.850	C_0
filamin A, alpha isoform 1	FLNA *	VHSPSGALEEcYVTEIDQDK[y]AVR	0.875	0.875	C_0
high density lipoprotein binding protein	HDLBP *	MD[y]VEINIDHK	0.867	0.867	C_0
homeodomain-interacting protein kinase 1 isoform 2	HIPK1 *	AVcST[y]LQSR	0.980	0.980	C_0
NKRF3 kinase family member	PEAK1	ASTDVAGQAVTINLVPTEEQAKP[y]R	0.523	0.523	C_0, C_6
phosphoinositide-3-kinase, regulatory subunit 2 (beta)	PIK3R2 *	SREYDQL[y]EEYTR	0.968	0.968	C_0
serine/threonine-protein kinase PRP4K	PRPF4B *	LcDFGSASHVADNDITP[y]LVSR	0.986	0.986	C_0
splicing factor, arginine/serine-rich 9	SRSF9 *	G[s]PHYFSPRPY	0.974	0.974	C_0
signal transducer and activator of transcription 3 isoform 1	STAT3	YcRPESQEHPADPGSAAPI[y]LK	0.903	0.903	C_0
tensin like C1 domain containing phosphatase isoform 1	TENC1 *	GPLDGSPI[y]AQVQRPPR	0.970	0.970	C_0
tensin	TNS1 *	DDGMEEVVGHTQGPDGSL[y]AK	0.958	0.958	C_0
vinculin isoform VCL	VCL *	SFLDSG[y]R	0.956	0.956	C_0
$C_1 = \{\mu \in R^4 : \mu_1 < \mu_2 < \mu_3 < \mu_4\}$					
docking protein 1	DOK1	VKEEGYELPYNPATDD[y]AVPPPR	0.069	0.506	C_1, C_2, C_3
SH2 containing inositol phosphatase isoform b	INPP5D	EKL[y]DFVK	0.000	0.721	C_1, C_2

Protein	pTyr	Sequence	$\hat{P}(Z_j = 0 Y)$	Maximum Posterior Probabilities	75% Credible Set
myelin protein zero-like 1 isoform a	MPZL1 *	SESVV[y]ADIR	0.077	0.586	C_1, C_5
$C_2 = \{\mu \in R^4 : \mu_1 < \mu_2 < \mu_3 > \mu_4\}$					
GRB2-associated binding protein 1 isoform a	GAB1 *	SSGSGSVADERVD[y]VVVDQQK	0.000	0.996	C_2
Wiskott-Aldrich syndrome gene-like protein	WASL *	VI[y]DIEIK	0.001	0.851	C_2
$C_3 = \{\mu \in R^4 : \mu_1 < \mu_2 > \mu_3 < \mu_4\}$					
mitogen-activated protein kinase 14 isoform 1	MAPK14 *	HTDDEMTG[y]VATR	0.000	0.771	C_3
$C_4 = \{\mu \in R^4 : \mu_1 < \mu_2 > \mu_3 > \mu_4\}$					
paxillin	PXN	VGEEHV[y]SFPNK	0.000	0.666	C_4, C_3
$C_5 = \{\mu \in R^4 : \mu_1 > \mu_2 < \mu_3 < \mu_4\}$					
mitogen-activated protein kinase 7 isoform 1	MAPK7	GLcTSPAEHQYFMTE[y]VATR	0.000	0.402	C_5, C_1
$C_6 = \{\mu \in R^4 : \mu_1 > \mu_2 < \mu_3 > \mu_4\}$					
TYRO protein tyrosine kinase binding protein isoform 1 precursor	TYROBP	ITE[TES]P[y]QEIQQR	0.044	0.946	C_6
$C_7 = \{\mu \in R^4 : \mu_1 > \mu_2 > \mu_3 < \mu_4\}$					
NCK adaptor protein 1	NCK1 *	L[y]IDLNMPAYVK	0.000	0.820	C_7
neural precursor cell expressed, developmentally down-regulated 9 isoform 1	NEDD9	EKD[y]DFFPPMR	0.010	0.463	C_7, C_3, C_8
$C_8 = \{\mu \in R^4 : \mu_1 > \mu_2 > \mu_3 > \mu_4\}$					
cyclin-dependent kinase-like 5	CDKL5 *	NLSEGNANYTE[y]VATR	0.020	0.668	C_8, C_6
ephrin receptor EphA2	EPHA2 *	VLEDDPEAT[y]ITTSQKPIR	0.000	0.844	C_8
FYN binding protein (FYN-120/130) isoform 2	FYB *	TTAVEID[y]DSLK	0.474	0.254	C_6, C_8, C_4
protein kinase C, delta	PRKCD *	RSDSASSEPVG[y]QGFKEK	0.000	0.419	C_8, C_4

* indicates that the pTyr site is classified to the same profile using both the ORI method and the Bayesian model.

Table 2

Posterior median and 95% CI for each model parameter

	Posterior Median	95% CI
σ	0.173	(0.164, 0.185)
μ_0	1.002	(0.929, 1.078)
σ_μ	0.278	(0.228, 0.347)
σ_{μ^*}	0.104	(0.080, 0.141)
σ_γ	0.017	(0.002, 0.076)
σ_η	0.003	(0.000, 0.017)
σ_δ	0.107	(0.076, 0.161)
ρ	0.131	(0.005, 0.504)
θ	0.487	(0.309, 0.673)

Simulation results from Section 4.1. 200 simulated datasets with $J = 35$ and $T = 4$ were used in both the simulation settings. In each simulation, 8 pTyr sites were assigned to the null profile C_0 and 27 pTyr sites were assigned to the other candidate profiles. The quantities reported in this table are averaged across 200 simulated datasets. The accuracy of the estimated trajectory is measured by the sum of squared errors (SSE), and a lower value indicates a higher accuracy.

Table 3

	$N=5$				$N=20$			
	Bayesian		ORI		Bayesian		ORI	
	MAP	MPW	MAP	MPW	MAP	MPW	MAP	MPW
Correct classification (overall)	94.1%	94.1%	68.3%	99.5%	99.5%	99.5%	71.2%	0.0%
Type I error rate: $Pr(\mu_j \in \cup_{k=1}^{C-1} \hat{C}_k \mu_j \in C_0)$	4.0%	4.0%	5.0%	0.6%	0.6%	0.6%	0.0%	0.4%
Type II error rate: $Pr(\mu_j \in \hat{C}_0 \mu_j \in \cup_{k=1}^{C-1} C_k)$	0.2%	0.2%	0.4%	0.0%	0.0%	0.0%	0.0%	38.5
Trajectory Accuracy: $\sum_{j=1}^J \sum_{t=1}^T (\hat{\mu}_{jt} - \mu_{jt})^2$	18.5	18.5	54.1	4.3	4.3	4.3	38.5	